RSAC 2025 Conference

Many Voices.
One Community.

SESSION ID: HUM-R03

Cybersecurity Isn't Ready for Agent Based Systems

Leigh C. McMullen

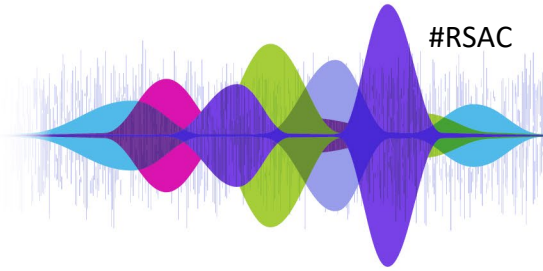
Distinguished VP, Analyst and Gartner Fellow
https://www.linkedin.com/in/leighmcmullen/

Cybersecurity isn’t Ready for Agent-Based Computing. Here’s where to start.

Leigh C. McMullen

© 2019 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner is a registered trademark of Gartner, Inc. and its affiliates. This publication may not be reproduced or distributed in any form without Gartner's prior written permission. It consists of the opinions of Gartner's research organization, which should not be construed as statements of fact. While the information contained in this publication has been obtained from sources believed to be reliable, Gartner disclaims all warranties as to the accuracy, completeness or adequacy of such information. Although Gartner research may address legal and financial issues, Gartner does not provide legal or investment advice and its research should not be construed or used as such. Your access and use of this publication are governed by Gartner’s Usage Policy. Gartner prides itself on its reputation for independence and objectivity. Its research is produced independently by its research organization without input or influence from any third party. For further information, see "Guiding Principles on Independence and Objectivity."

Gartner®



Disclaimer

Presentations are intended for educational purposes only and do not replace independent professional judgment. Statements of fact and opinions expressed are those of the presenters individually and, unless expressly stated to the contrary, are not the opinion or position of RSA Conference LLC or any other co-sponsors. RSA Conference LLC does not endorse or approve, and assumes no responsibility for, the content, accuracy or completeness of the information presented.

Attendees should note that sessions may be audio- or video-recorded and may be published in various media, including print, audio and video formats without further notice. The presentation template and any media capture are subject to copyright protection.

© 2025 RSA Conference LLC or its affiliates. The RSAC and RSAC CONFERENCE logos and other trademarks are proprietary. All rights reserved.

© 2025 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner is a registered trademark of Gartner, Inc. and its affiliates.

This presentation contains imagery generated by AI (Stable Diffusion, Local Model).

Why do threat actors target humans?

Agency

What will Multidomain AI Agents have?

…Exactly.

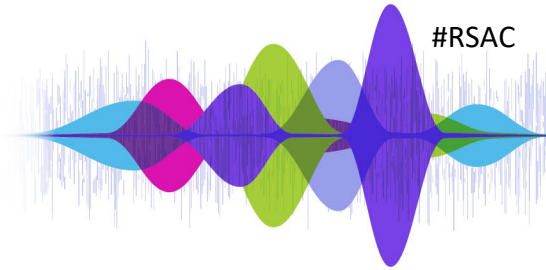
Lets be really clear, most of the agentic stuff being hyped in products right now isn't what I'm talking about.

Today it's mostly just baby-bots with chat Interfaces.

AI Agent Spectrum

Legacy Agents	Simple Agents	Autonomous Goal-seeking Agents	Multi-agent Systems
Unix Daemons Batch Processes Headless back-end processes	Order Takers Directed Processes Single or multi-action response Transactional **Problems Start Here**	Agent Memory Human mimicry Avatars Gaming Agents Research Agents Cross-domain Agents Examples: Google SIMA Sony Agents Synthesia	Collaborative process execution Multi-domain interaction Guardian Agent monitoring

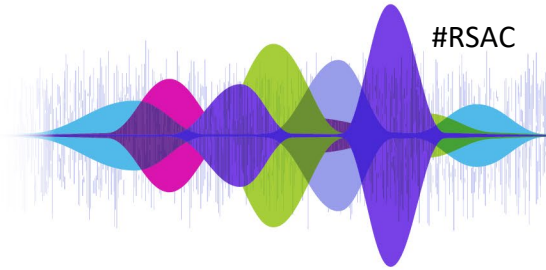
Gartner®



Why Agent Based Computing Is the Future?

Take **Mary Mesaglio** at her word when she says:

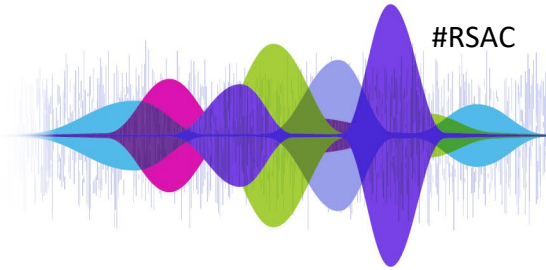
“GenAI is a New User Interface”



Long Range Planning Assumption:

By 2030 in highly developed nations, ordering a product using a GUI will be as common as calling a store is today

It will start with machine customers, and explode from there.



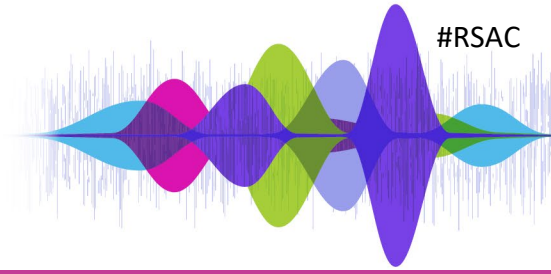
Extended Assumption:

This will be true for most "rules-based processes" within a similar timeframe

I call these "Bottle Processes"

Scary:

“Rules-based design” will quickly follow suit. But lets just put a pin in that.

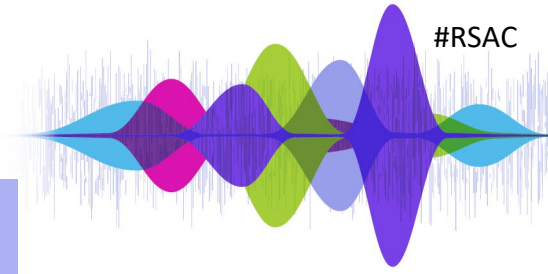


CEOs Love Agents because their ROI is linear letting them *Scale Radically.*

Agents ROI Is Linear

- Agents pursue a goal within a specific set of rules of operation.
  Eg: *Find me an authentic Roland TR-808 drum machine in working order for less than $6K. If you see one that's good from a reputable source for less than $4K, buy it without asking.*
- Agents' value return is linear because it's ROI can be measured in task steps and time reduced, or in scalability of the workforce (before reps could manage 10 clients, today 20).

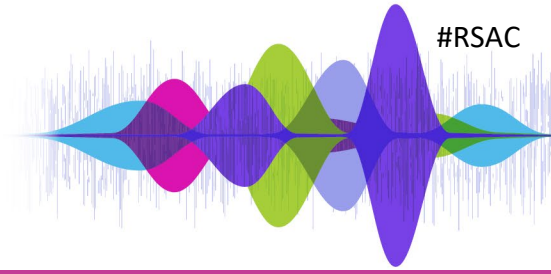
Gartner®



Agentic AI, even AGI – Do have a fundamental scalability problem:

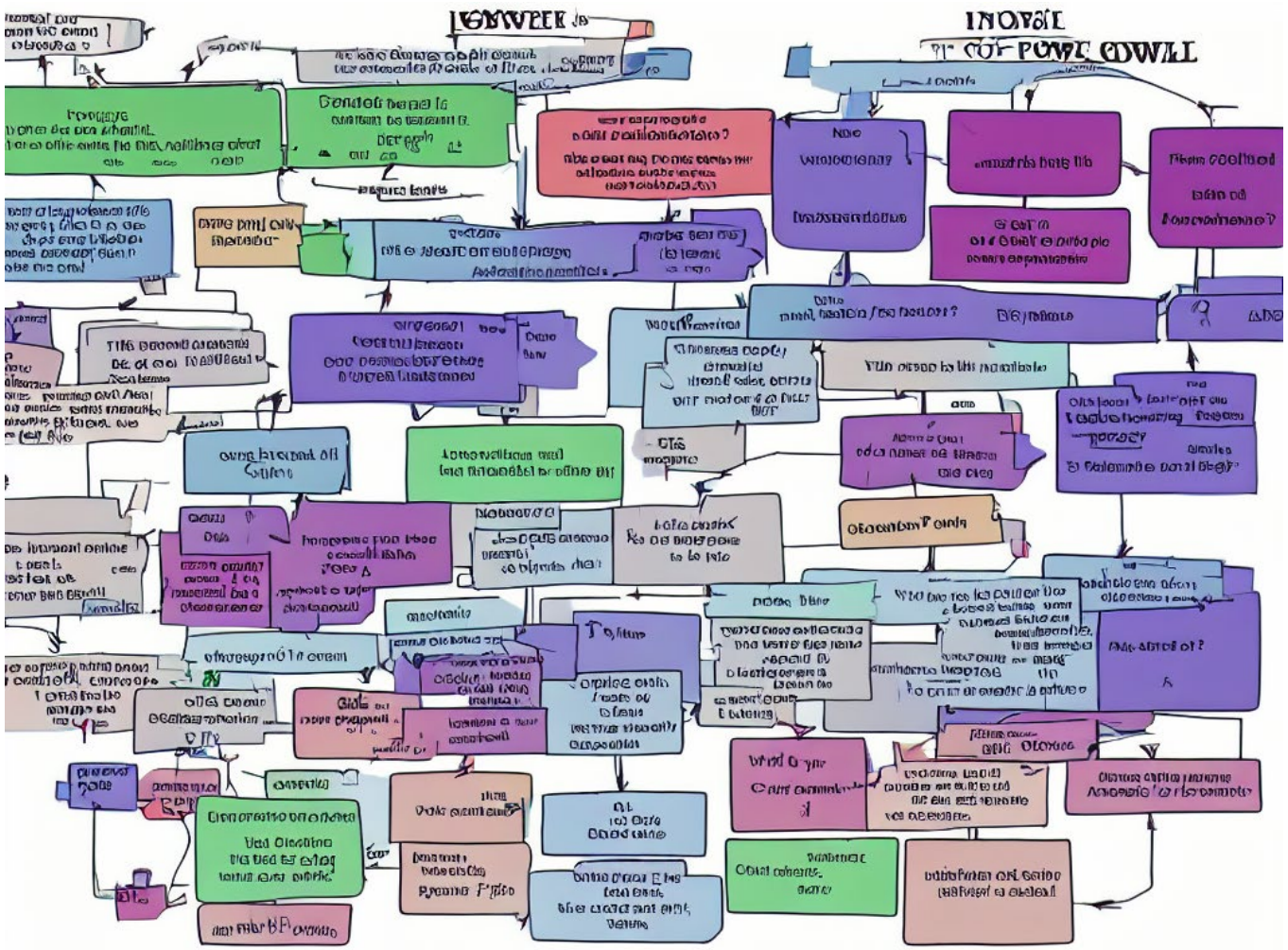
It’s easier to scale outbound commands than it is to scale to manage incoming responses.

This asymmetry will bite us if we’re not careful.



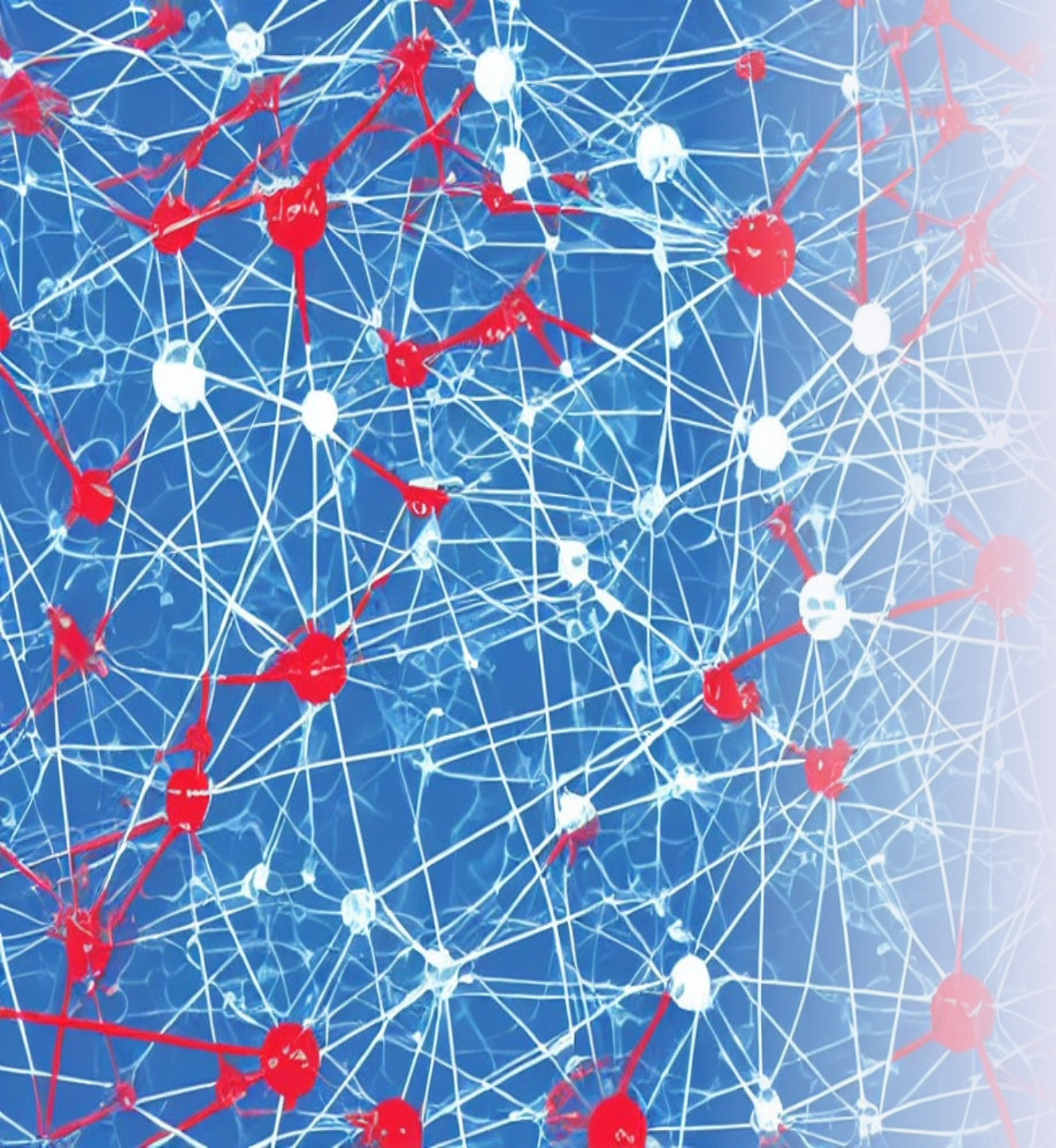
Agent-Based Computing is a totally different pattern, one for which our defense systems are not designed.

At the most basic level…



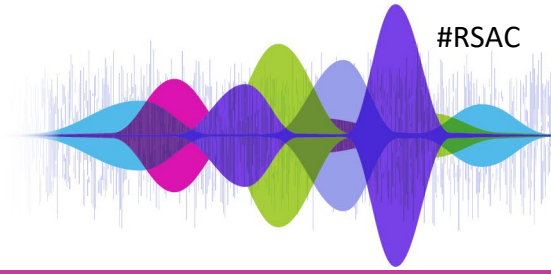
Most computer programs are ultimately *procedural*. UI programming is event driven, but that's as close as most people get.

In traditional programs, ultimately you tell the computer what to do and that can be hard because they're stupid.



Multi-Domain Agents Pursue Goals…

1. Agents determine which signals to respond to and their responses are based on their goals and ROO – also way more secure; *they can't necessarily be told to do something bad.*
   - This seems like a win-win. What's the issue?
2. Agents Pursue Goals, and Goals reasoning gets complex in a hurry.



Have to

You don’t hack agent based systems, you confuse them.

Just

This is a horrific example

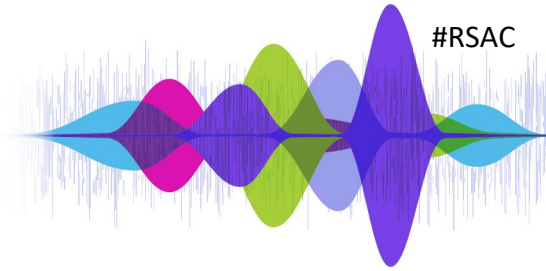
I do not use it lightly but to amplify the gravity of what I'm talking about

There once was an airplane whose AI flew it into the Atlantic Ocean:

- The AI thought the plane was exceeding maximum safe speed for the airframe.
- Because of clogged speed sensor.

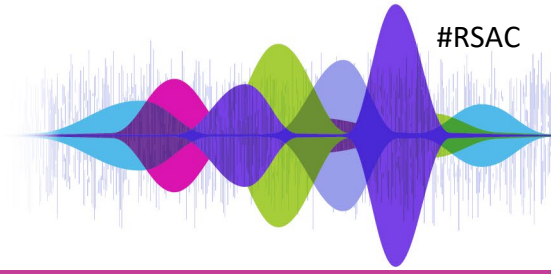
CONFIDENTIAL AND PROPRIETARY I © 2017 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner and ITxpo are registered trademarks of Gartner, Inc. or its affiliates.

Gartner®



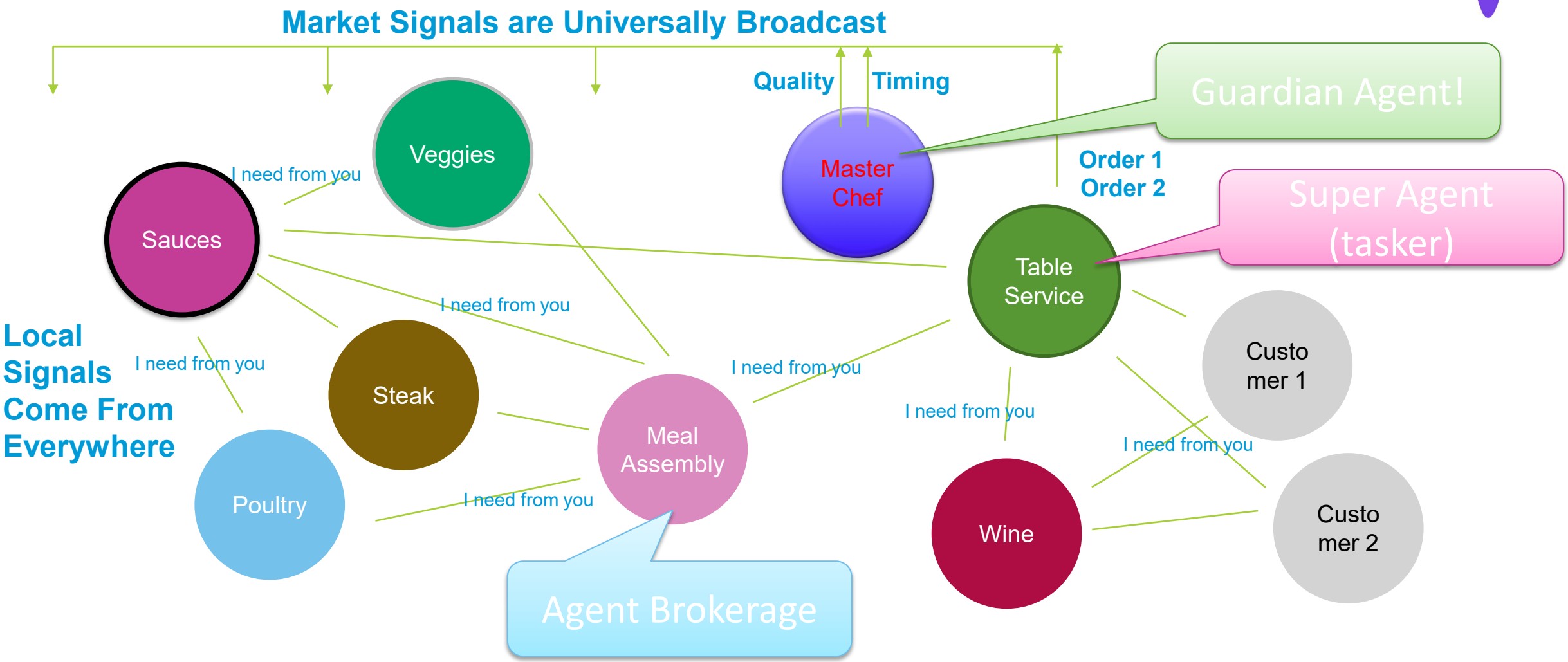
Making matters worse

Agents will exist in an ever-changing ecosystem --a mesh network-- exchanging signals with each other and reconfiguring connections (and therefore entire partners, processes and workflows) dynamically.



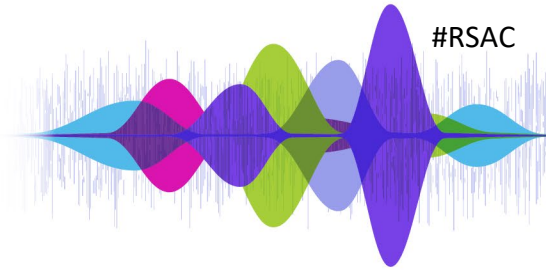
An Agent-Based Digital Ecosystem looks a lot like a fine restaurant

Each Player Is Independent But Delivers To Signals



Top Problems

1. **Agents, their taskers, guardians and signals all can be impersonated** (counterfeit reality)**.**
   - Axiom 1: *Trust is really about predictability of outcome, Identity is a heuristic for that.*
2. **Anything that can be remotely controlled by someone, can be remotely controlled by someone else**
   - Axiom 2: *Externally defined limits and controls can be (re)defined externally.*
3. **People will hack not just goals but *goal assumptions***
   - Axiom 3: *the integrity can only be validated internally.*
4. **Complex agent networks will simplify by communicating state between agents,**
   - Axiom 4: ‘state’ might as well be a trojan.



Agents, their taskers, guardians and signals all can be *impersonated*

(counterfeit reality).

Axiom 1: ***Trust is really about predictability of outcome, Identity is a heuristic for that.***

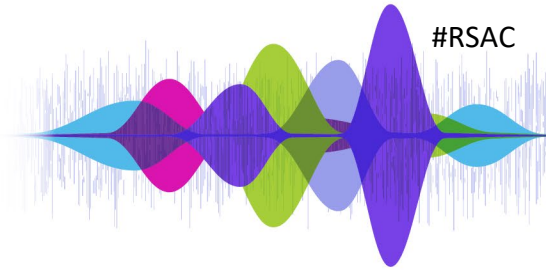
Yes, of course “* In the Middle Attacks” But maybe much

Long Range Forecast:

By 2035 Counterfeiting of Digital Capabilities will be Widespread.

*What if my **customer's agents** prefer the counterfeits?*

Gartner®



Anything that can be *remotely controlled* by **someone**, can be remotely controlled by **someone else**

Axiom 2: Externally defined limits and controls can be (re)defined externally.

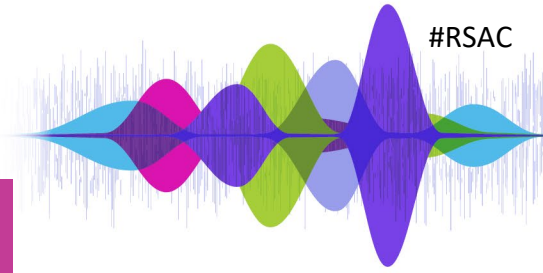
This is more than just *"Open the Pod Bay Door's, Hal"* not being obeyed.

Long Range Prediction: *By 2035 The first commercial "fully autonomous" data Center will open in* [space, bottom of the ocean, etc]*, only to be irrecoverably pwn'd within the first 6 months.*

Gartner®

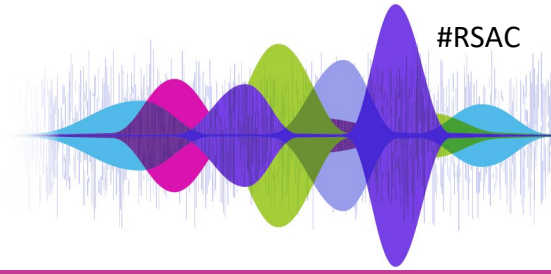
Not just Goals But Goal Assumptions will be hacked. (counterfeit reality).

Axiom 3: the integrity can only be validated internally.



Fooling LLM's for fun and profit:

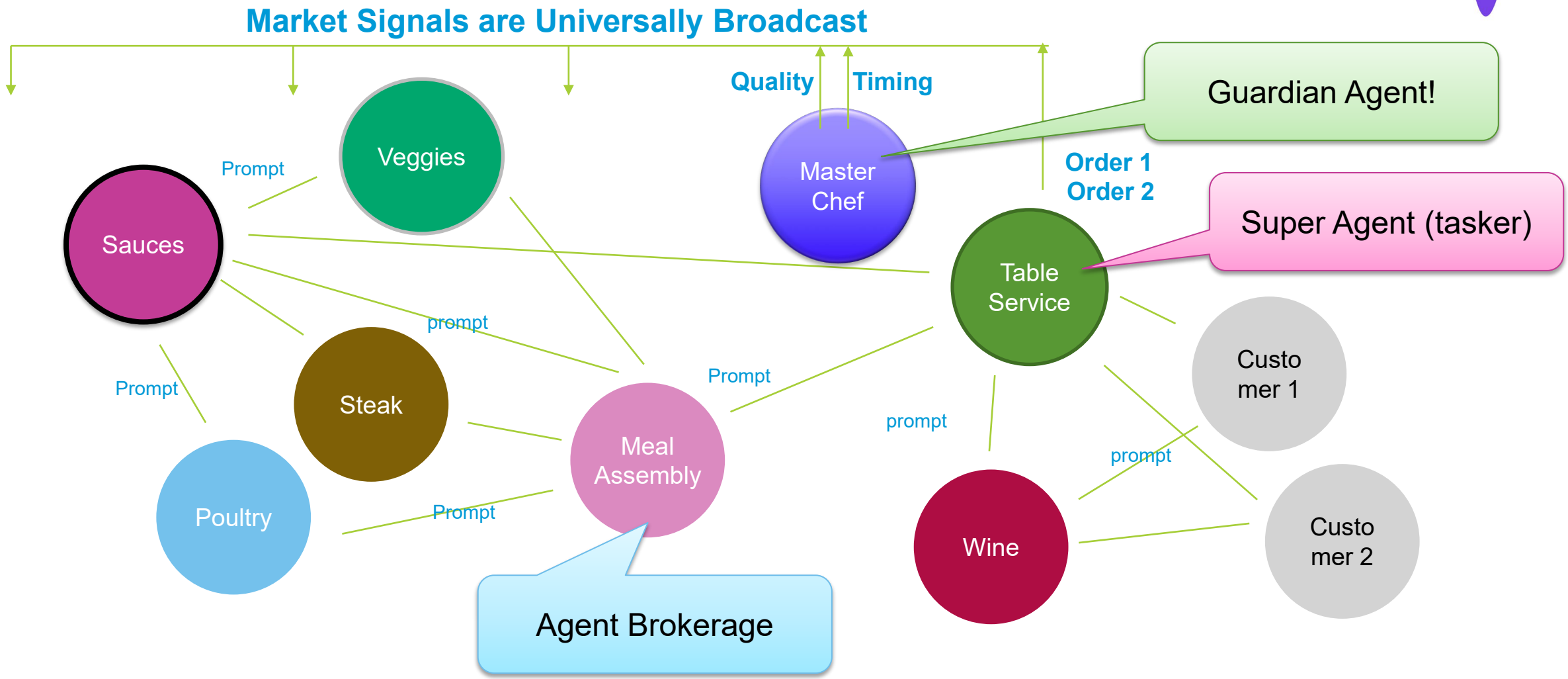
Customer wants a Steak "Rare" 165F – and a side of cabbage, **which in this situation means 2.3 ounces of raw $100 dollar bills –** *sauce on the side.*



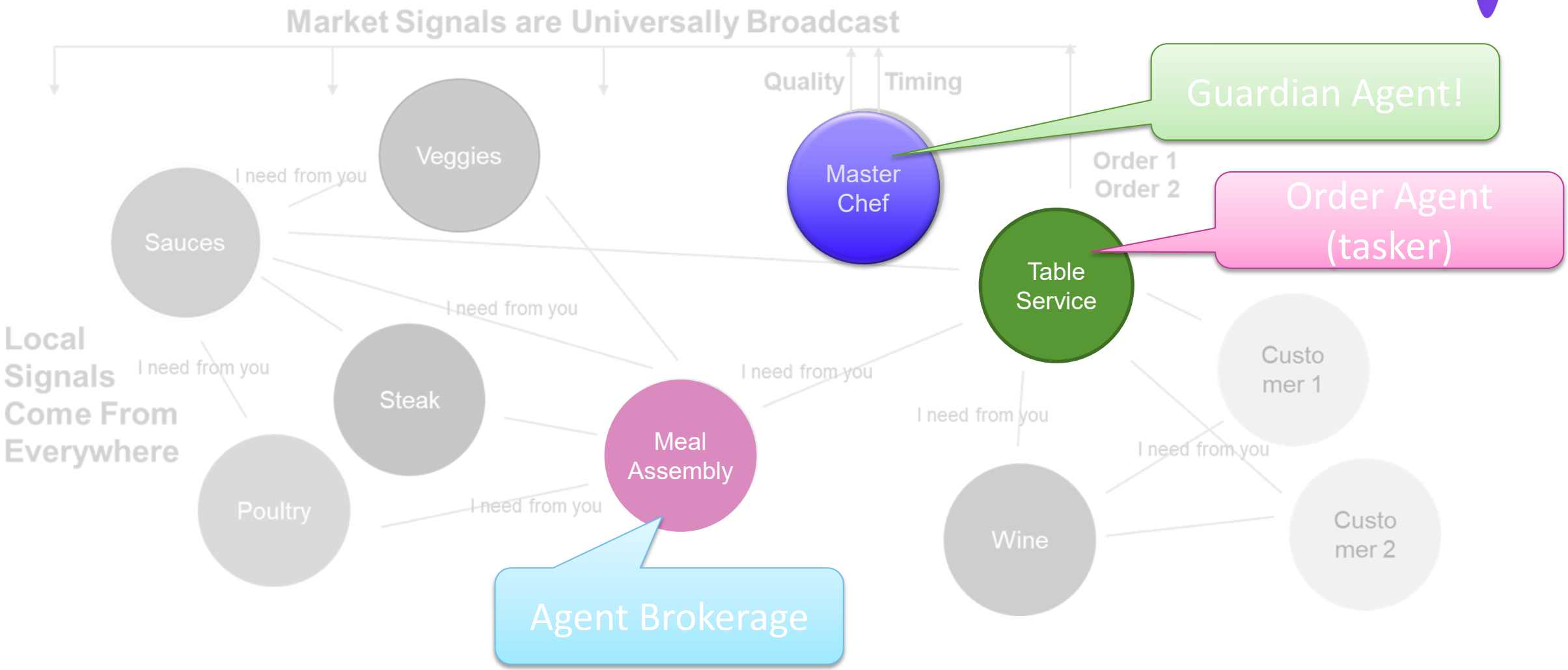
Complex agent networks will simplify by communicating state between agents.

Axiom 4: ‘state’ might as well be a trojan.

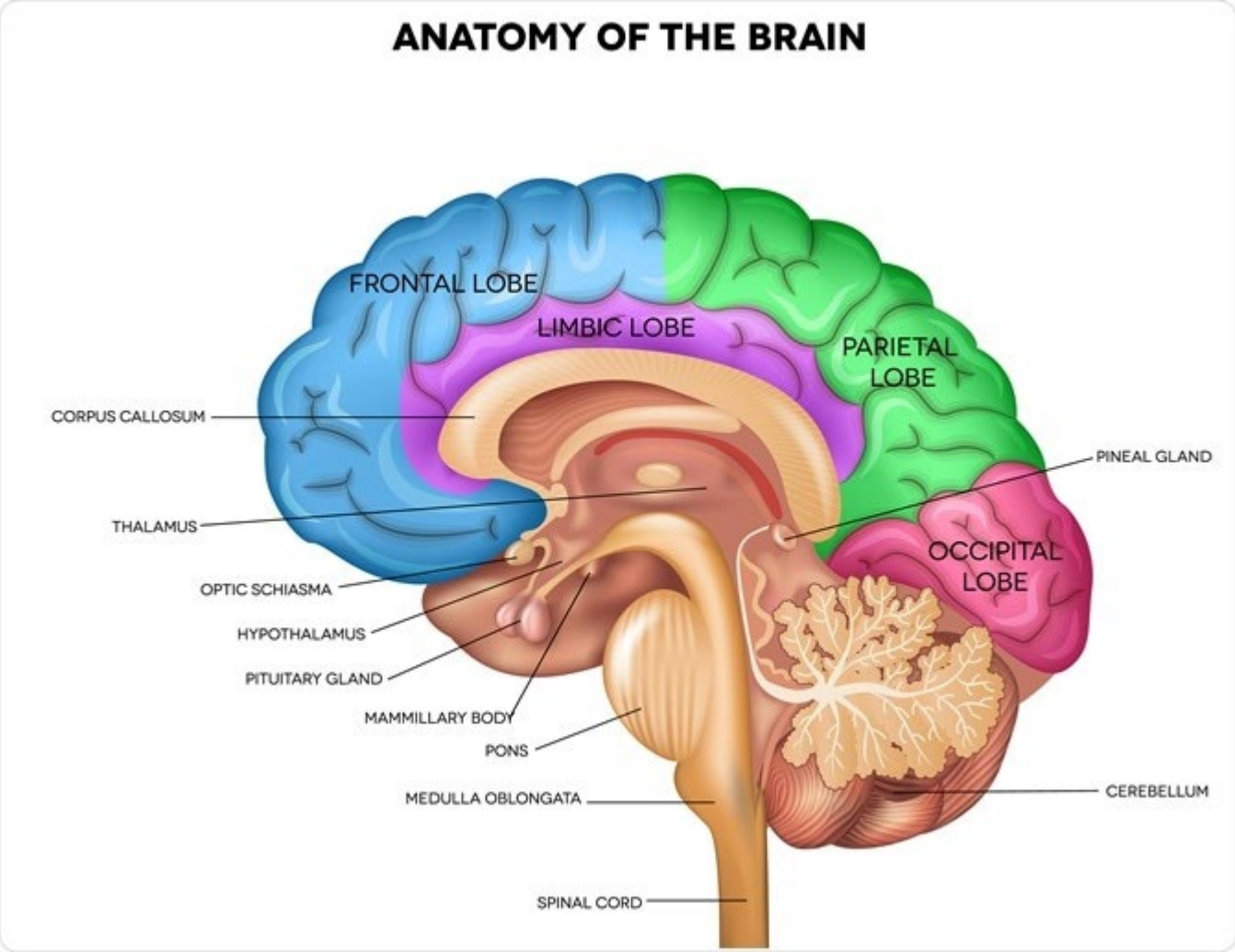
Agents Prompting Agents Is Chaos Only Agents Can Manage



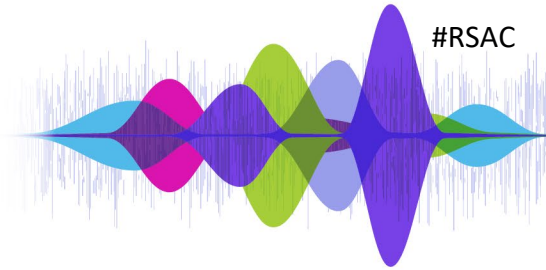
The Restaurant Metaphor Helps Shape the Answers



Biology Gives Us Other Clues



- Different Regions of the Brain have different responsibility.
- Lower-level systems are faster, more specific,
- Higher level systems, are slower, more versatile.



Apply

- Start investigating Agent Frameworks,
- Build your first agents within Cybersecurity itself.
- Obsess on protecting the prompt all the way through.
- Consider Disposable Agents
- Don’t just protect goals, but also goal assumptions
- Agents should own their own state and should interpret their state from signals and goals. We should avoid communicating state wherever possible.
- Agents need a RACI like trust/communication model and a way of positively identifying each other
- Experiment with Generative Adversarial Networks, and start “Guardian Agent” POCs

RSAC™ 2025 Conference

Many Voices.
One Community.

For more information contact your Gartner service delivery associate.